

Simple Search Engine Model: Adaptive Properties

Mahyuddin K. M. Nasution

Information Technology Department,
Fakultas Ilmu Komputer dan Teknologi Informasi
Universitas Sumatera Utara, Padang Bulan, Medan 20155, Sumatera Utara, Indonesia
`mahyunst@yahoo.com, mahyuddin@usu.ac.id`

Abstract. In this paper we study the relationship between query and search engine by exploring the adaptive properties based on a simple search engine. We used set theory and utilized the words and terms for defining singleton space of event in a search engine model, and then provided the inclusion between one singleton to another.

Keywords: singleton space, information retrieval, search term, query.

1 Introduction

A search engine on the World Wide Web, in brief we called it as Web, is extensively important to help users to find relevant information. The search engines have some features for servicing the tasks and subtasks that directly or indirectly uses the techniques such as indexing, filters, hub, page rank, hits, and etc [1], but to access any information in Web the users need the formulating a query about the required information. In this case, the query has become the leading paradigm to find the information, whereby the information retrieval (IR) is concerned with answering information need as accurately as possible. However, the users lack understand a formulae of query. Moreover, almost all of search of engines is not provide any function to find the special cases such as entity or actors. Therefore, the major challenge in information access is to provide the riched and trusted information. This paper is aimed at generating some adaptive properties of relation between an search engine and a query.

2 Basic Concept and Motivation

Let objects (entities or attributes) can be given literally, like the literal text of "Social Network", then all meaning of objects based on words is represented by the literal objects itself. To realize it, first we define formally that a word w is the basic unit of discrete data, defined to be an item from a vocabulary indexed by $\{1, \dots, K\}$, where $w_k = 1$ if $k \in K$, and $w_k = 0$ otherwise [2]. Then, we define some instances related to words.

Definition 1. A term t_x consists of at least one or a set of words in a pattern, or $t_k = (w_1 w_2 \dots w_l)$, $l \leq k$, k is a number of parameters representing word w , l is the number of tokens (vocabularies) in t_k , $|t_k| = k$ is size of t_k . ■

We define a simple search engine as follows.

Definition 2. Let a set of web pages indexed by search engine be Ω , i.e., a set contains ordered pair of the terms t_{k_i} and the web pages ω_{k_j} , or (t_{k_i}, ω_{k_j}) , $i = 1, \dots, I$, $j = 1, \dots, J$. The relation table that consists of two columns t_k and ω_k is a representation of (t_{k_i}, ω_{k_j}) where $\Omega_k = \{(t_k, \omega_k)_{ij}\} \subset \Omega$ or $\Omega_k = \{\omega_{k_1}, \dots, \omega_{k_j}\}$. The cardinality of Ω is denoted by $|\Omega|$. ■

In Definition 2, we assume that Ω is made of a set of index of terms t_{k_i} , we will call it as a space of term. So, the web pages and queries are represented as vectors in Ω is also a space of event, whereby the semantics of this space is that of a multidimensional space. Therefore, a term t_k is represented as a vector of web pages, i.e., the meaning of a term to be $\omega_k \in \Omega$ in which t_k occurs. Let q is a query, then $t_k \in q$, for $t_k = (w_1 w_2 \dots w_k)$. In logical implication, a web page is relevant to a query if it implies the query, that is if $\omega \Rightarrow q$ is true or $\omega \Rightarrow t_k$ is true $\forall \omega \in \Omega$: $(\omega \Rightarrow t_k) = 1$ [3], but for $2^k - 2$ others of $\{\{t_k^{2^k-2}\} \subset \{w_1, w_2, \dots, w_k\} = t_k\}$, also $\omega \Rightarrow q$ is true $\forall \{t_k^{2^k-2}\} \neq \emptyset$. Thus, the degree of $\omega \Rightarrow q$ measured by $P(\omega \Rightarrow q)$, and probability t_x in power subsets of $\{w_1, w_2, \dots, w_k\}$,

$$P(t_k) = \frac{1}{2^k - 1}, t_k = (w_1 w_2 \dots w_k). \quad (1)$$

Therefore there are an uniform mass probability function for Ω ,

$$P : \Omega \rightarrow [0, 1] \quad (2)$$

where $\sum_{\Omega} P(\omega) = 1$.

Definition 3. Let t_x is a search term, and $t_x \in \mathcal{S}$ where \mathcal{S} is a set of singleton search term of search engine. A vector space $\Omega_x \subseteq \Omega$ is a singleton search engine event (singleton space of event) of web pages that contain an occurrence of $t_x \in \omega_x$. The cardinality of Ω_x is denoted by $|\Omega_x|$. ■

In the singleton space of event, $\Omega_x \subseteq \Omega$ if $\omega \Rightarrow t_x$ is true, or

$$\Omega_x(t_x) = \begin{cases} 1 & \text{if } t_x \text{ is true at } \omega \in \Omega, \\ 0 & \text{otherwise} \end{cases} \quad (3)$$

and the cardinality of Ω_x be $|\Omega_x| = \sum_{\Omega} (\Omega_x(t_x) = 1)$. This means that every web page that is indexed by search engine contains at least one occurrence of a search term, then we can measure its degree of uncertainty of $\omega \Rightarrow t_x$ on $\omega \Rightarrow q$ by

$$P(\Omega_x) = P(\Omega_x(t_x) = 1) = \frac{\sum_{\Omega} (\Omega_x(t_x) = 1)}{|\Omega|} = \frac{|\Omega_x|}{|\Omega|} \quad (4)$$

For example, a search term is a person name: $x = \text{"Mahyuddin Khairuddin Matyuso Nasution"}$, then $\{t_x\} = \{w_1, w_2, w_3, w_4\} = \{\text{"Mahyuddin"}, \text{"Khairuddin"}, \text{"Matyuso"}, \text{"Nasution"}\}$. At the time of doing the experiment, a Yahoo! search for "Mahyuddin Khairuddin Matyuso Nasution" returned $|\Omega_x| = ?$ hits or $|\Omega_x| = 3,440$ for "Mahyuddin K. M. Nasution", and the number of hits of search for $w_{i=1,2,3,4}$ are in $\{54,300; 3,187,000; 0; 275,000\}$. The vector space of t_x is of $\{t_k^{2^k-1}\} = \{\{w_1\}, \{w_2\}, \{w_3\}, \{w_4\}, \{w_1, w_2\}, \{w_1, w_3\}, \{w_1, w_4\}, \{w_2, w_3\}, \{w_2, w_4\}, \{w_3, w_4\}, \{w_1, w_2, w_3\}, \{w_1, w_2, w_4\}, \{w_2, w_3, w_4\}, \{w_1, w_2, w_3, w_4\}\}$. We have also $|\Omega_{x_p}| = 55$ from Yahoo search engine for t_x with its pattern as a meaning core of Ω_x ,

$$|\Omega_{x_p}| = \sum_{\Omega} (\omega_x \Rightarrow t_x) \leq |\Omega_x|, \quad (5)$$

where $\sum_{\Omega} (\omega_x \Rightarrow t_x)$ is the number of web pages containing t_x with the pattern exactly. The singleton space of event captures in a particular sense all background knowledge about the search terms concerned available on the Web, geometrically this is a representation of meaning semantically.

Similarly, for two search terms t_x and t_y in the different queries, we have

$$\Omega_x \cap \Omega_y = (\Omega_x(t_y) = 0) \wedge (\Omega_y(t_x) = 0) = \emptyset \quad (6)$$

i.e. any two singleton spaces of event are independent.

Problem 1. Let t_x and t_y are two different search terms, $t_x \neq t_y$. Let Ω_x and Ω_y are the singleton search engine events of t_x and t_y , respectively, and $|t_y| < |t_x|$ or $\forall w_i \in t_y, w_i \in t_x, \exists w_j \in t_x, w_j \notin t_y$, then

$$|\Omega_x| \stackrel{?}{=} |\Omega_x| + |\Omega_y| \quad (7)$$

where $\Omega_x, \Omega_y \subseteq \Omega$.

Problem 1 is a property of relation between any search engine and any query in a heterogeneous environment such as Web, and the information about any object to be scattered in various places. So in almost all measurements the bias exist.

3 The Adaptive Properties in Search Engine

Numerous studies of natural language processing (NLP) and Semantic Web utilize a search engine, mainly to obtain a set of documents that include a given query and to get statistical information about an object such as hit count of entity name, but to bring the NLP and Semantic Web to life such as the information processing services provide the knowledge, for example: ontology construction, knowledge extraction, question answering, and other purposes [4] needs more effort.

Some properties we will derive to learn how to get the efficient ways to access and extract information from web. The purpose of this construction is to eliminate the bias by developing the adaptive model of relation between a search engine and the search terms.

Lemma 1. Let t_x and t_y are search term. If $t_x \neq t_y$, $t_x \cap t_y \neq \emptyset$ and $|t_y| < |t_x|$, then singleton search engine event of t_x and t_y is $\Omega_x = \Omega_x \cup \Omega_y$ or

$$|\Omega_x| = |\Omega_x| + |\Omega_y|, \quad (8)$$

where $\Omega_x, \Omega_y \subseteq \Omega$.

Proof. For all search terms t_x and t_y where $t_x \neq t_y$, $t_x \cap t_y \neq \emptyset$ and $|t_y| < |t_x|$, by Definition 1 and Definition 2 we have $\forall w_y \in t_y, w_y \in t_x, \exists w_x \in t_x, w_x \notin t_y \Rightarrow \forall w_y \in \omega_y, w_y \in \omega_x, \exists w_x \in \omega_x, w_x \notin \omega_y$ such that

$$t_x \cap t_y = t_y \text{ and } t_x \cup t_y = t_x \quad (9)$$

and

$$\omega_x \cap \omega_y = \omega_y \text{ and } \omega_x \cup \omega_y = \omega_x. \quad (10)$$

By Eq. (6), clear that $\Omega_x \neq \Omega_y$ and $|\Omega_x \cap \Omega_y| = 0$, then we have

$$|\Omega_x \cup \Omega_y| = |\Omega_x| + |\Omega_y|. \quad (11)$$

Let $\Omega_x = \{(t_x, \omega_x)\}$, based on meaning Eq. (9) and Eq. (10), we have $\Omega_x = \{(t_x, \omega_x)\} = \{(t_x \cup t_y, \omega_x \cup \omega_y)\} = \{(t_x, \omega_x) \cup (t_y, \omega_y)\} = \{(t_x, \omega_x)\} \cup \{(t_y, \omega_y)\} = \Omega_x \cup \Omega_y$. Therefore based on Eq. (11) the Eq. (7) in Problem 1 be $|\Omega_x| = |\Omega_x| + |\Omega_y|$. ■

Proposition 1. Let t_z, \dots, t_y, t_x are search terms, where $t_z \neq \dots \neq t_y \neq t_x$ and $|t_z| < \dots < |t_y| < |t_x|$, then $\Omega_x = \Omega_x \cup \Omega_y$ holds recursively or $|\Omega_x| = |\Omega_x| + |\Omega_y|$, $\Omega_x, \Omega_y \subseteq \Omega$.

Proof. By the Lemma 1 and an assumption that $|t_z| < \dots < |t_y| < |t_x|$, we obtain $|t_z| < |t_{z1}| \Rightarrow |\Omega_{z1}| = |\Omega_{z1}| + |\Omega_z|$, $|t_{z1}| < |t_{z2}| \Rightarrow |\Omega_{z2}| = |\Omega_{z2}| + |\Omega_{z1}|$, \dots , $|t_y| < |t_x| \Rightarrow |\Omega_x| = |\Omega_x| + |\Omega_y|$. Because of the inter-independence in the queries such as Eq. (6), we obtain $\Omega_x \cap \Omega_y = \emptyset, \dots, \Omega_{z1} \cap \Omega_z = \emptyset$, and $\Omega_x \cup \Omega_y$ belonging to Ω_x , then

$$\begin{aligned} |\Omega_x| &= |\Omega_x \cup \Omega_y| \\ &= |\Omega_x| + |\Omega_y| \\ &= |\Omega_x| + |\Omega_y \cup \dots| \\ &= |\Omega_x| + |\Omega_y| + \dots \\ &= |\Omega_x| + |\Omega_y| + |\dots \cup \Omega_z| \\ &= |\Omega_x| + |\Omega_y| + \dots + |\Omega_z| \end{aligned}$$

or $|\Omega_x| = |\Omega_x| + |\Omega_y|$ be recursive, where $|\Omega_y| + \dots + |\Omega_z|$ is a part of $|\Omega_x|$. ■

Lemma 2. If $t_y \neq t_z$ and $t_y \cap t_z = \emptyset$, then $|\Omega_y \cap \Omega_z| = 0$ and $|\Omega_y \cup \Omega_z| = |\Omega_y| + |\Omega_z|$.

Proof. For all search terms t_y and t_z where $t_y \neq t_z$ and $t_y \cap t_z = \emptyset$, by Definition 1 and Definition 2 we obtain $\forall w_y \in t_y, w_y \notin t_z \wedge \forall w_z \in t_z, w_z \notin t_y \Rightarrow \forall w_y \in \omega_y, w_y \notin \omega_z \wedge \forall w_z \in \omega_z, w_z \notin \omega_y$ such that

$$t_z \cap t_y = \emptyset \vee t_z \cup t_y = t_y \cup t_z \quad (12)$$

and

$$\omega_y \cap \omega_z = \emptyset \vee \omega_y \cup \omega_z = \omega_z \cup \omega_y \quad (13)$$

Let $\Omega_y = \{(t_y, \omega_y)\}$ and $\Omega_z = \{(t_z, \omega_z)\}$ are two independent events from the queries, based on Eq. (6) we obtain $\Omega_y \cap \Omega_z = \emptyset$ and

$$|\Omega_y \cap \Omega_z| = 0 \quad (14)$$

and by combining the meaning of (12), (13), (14), and $\{(t_y, \omega_y)\} \cup \{(t_z, \omega_z)\} = \Omega_y \cup \Omega_z$ and we can conclude that $|\Omega_y \cup \Omega_z| = |\Omega_y| + |\Omega_z|$. ■

Lemma 2 expresses that Eq. (7) in Problem 1 be $|\Omega_x| \neq |\Omega_x| + |\Omega_y|$ or $|\Omega_x \cup \Omega_y| = |\Omega_x| + |\Omega_y|$.

Proposition 2. Let $\Omega_x \cap \Omega_y = \emptyset$ and $\Omega_a \cap \Omega_b = \emptyset$. If $|\Omega_x| = |\Omega_x| + |\Omega_a|$ and $|\Omega_y| = |\Omega_y| + |\Omega_b|$, then $|\Omega_x \cap \Omega_y| \geq 0$.

Proof. This is a direct consequence of Lemma 1 and Lemma 2. ■

Lemma 3. Let t_x and t_z are search terms. If $t_x \neq t_z$, $t_x \cap t_z = \emptyset$, and $\omega_x \cap \omega_z \neq \emptyset$, then $|\Omega_x| = |\Omega_z|$, $\Omega_x, \Omega_z \subseteq \Omega$.

Proof. For all search terms t_x and t_z where $t_x \neq t_z$, $t_x \cap t_z = \emptyset$ and $\omega_x \cap \omega_z \neq \emptyset$, by Definition 1 and Definition 2 we obtain $\forall w_x \in t_x, w_x \notin t_z$, and $\forall w_z \in t_z, w_z \notin t_x$ then

$$t_x \cap t_z = \emptyset \vee t_x \cup t_z = t_z \cup t_x, \quad (15)$$

but $\forall w_x \in \omega_x, w_x \in \omega_z$ and $\forall w_z \in \omega_z, w_z \in \omega_x$ then

$$\omega_x \cap \omega_z = \omega_x = \omega_z, \omega_x \cup \omega_z = \omega_z \cup \omega_x = \omega_x = \omega_z. \quad (16)$$

For $\Omega_x = \{(t_x, \omega_x)\}$ and $\Omega_z = \{(t_z, \omega_z)\}$ we have $\Omega_x \cap \Omega_z = \{(t_x, \omega_x)\} \cap \{(t_z, \omega_z)\} = \{(t_x, \omega_z)\} \cap \{(t_z, \omega_z)\}$, and because $t_z \in \omega_z$ the intersection of $t_x \cap t_z$ must be $\{(t_x, \omega_z)\} \cap \{(t_z, \omega_z)\} = \{(t_z, \omega_z)\} \cap \{(t_z, \omega_z)\}$ or $\Omega_x \cap \Omega_z = \Omega_z \cap \Omega_z$ or $\Omega_x \cap \Omega_z = \Omega_z$. Similarly, $\Omega_x \cap \Omega_z = \Omega_x$. Thus $|\Omega_x| = |\Omega_z|$. ■

This Lemma explains that Eq. (7) in Problem 1 be $|\Omega_x| = |\Omega_y|$ if and only if $t_x \neq t_y$ but $t_x, t_y \in \omega_x \wedge t_x, t_y \in \omega_y$. In other word, based on combining (15) and (16) $\Omega_x = \{(t_x, \omega_x)\} = \{(t_x, \omega_x \cup \omega_y)\} = \{(t_x, \omega_x) \cup (t_x, \omega_y)\} = \{(t_y, \omega_x) \cup (t_y, \omega_y)\} = \{(t_y, \omega_x \cup \omega_y)\} = \{(t_y, \omega_y)\} = \Omega_y$. This shows that the search terms may be different but they come from same web pages, and in this case they take the same meaning from web.

4 Conclusions and Future Work

Studying to properties of relation between query and search engine gave the understanding about the semantic representation statistically for object in literal text. Our near future work is to generate some properties of search engine for doubleton.

References

1. W. B. Croft, D. Metzler, and T. Strohman. *Search Engines Information Retrieval in Practice*. Addison Wesley. 2010.
2. D. M. Blei, A. Y. Ng, and M. J. Jordan. Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3: 993-1022, 2003.
3. M. K. M. Nasution and S. A. Noah. Information retrieval model: A social network extraction perspective. In *IEEE Proc. of CAMP 2012*: 322-326, 2012.
4. P. Cimiano, S. Handschuh, and S. Staab. Towards the self-annotating web. In *Proc. WWW 2004*: 462-471, 2004.